# GESTIÓN DE RIESGO DE MODELOS DE CRÉDITO CON *MACHINE LEARNING*: DESARROLLO, EXPLICABILIDAD Y MONITOREO AUTOMATIZADO

Eduardo Rafael Jauregui Romero
https://orcid.org/0009-0008-0248-2601
eduardo.jauregui1@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos, Perú

Recibido: 26 de noviembre del 2025 / Aceptado: 23 de junio del 2026
doi: https://doi.org/10.26439/interfases2026.n023.8406

RESUMEN. La adopción de modelos de aprendizaje automático en el sector bancario peruano enfrenta el desafío crítico de garantizar su estabilidad y transparencia en entornos productivos, conforme a las exigencias de la regulación vigente sobre gestión de riesgo de modelos. Este estudio presenta no solo el desarrollo de un modelo de calificación crediticia basado en el algoritmo HistGradientBoosting optimizado, sino, principalmente, la implementación de una arquitectura integral de monitoreo automatizado desplegada en servicios de computación en la nube. Utilizando una muestra histórica de 50 206 registros, se diseñó un flujo de trabajo que integra ingeniería de características y validación cruzada y que ha logrado un indicador de Gini del 67,4 %. La principal contribución radica en el despliegue de una plataforma web interactiva que permite la vigilancia en tiempo real de métricas de desempeño y estabilidad. A través de este sistema, se evaluó la consistencia del modelo mediante el índice de estabilidad poblacional y se obtuvo un valor de 0,4 %, lo que confirma la ausencia de degradación predictiva o de cambios estructurales en las variables a lo largo del tiempo. Este enfoque demuestra cómo la integración de técnicas de *machine learning operations* (MLOps) y herramientas de visualización avanzada facilita la auditoría continua y la detección temprana de desviaciones, lo que supera las limitaciones de los reportes estáticos tradicionales.

PALABRAS CLAVE: riesgo de crédito / *machine learning* / monitoreo de modelos / gestión de riesgo de modelos

## CREDIT MODEL RISK MANAGEMENT WITH MACHINE LEARNING: DEVELOPMENT, EXPLAINABILITY, AND AUTOMATED MONITORING

ABSTRACT. The adoption of machine learning models in the Peruvian banking sector faces the critical challenge of ensuring stability and transparency in production environments, in

compliance with current model risk management regulations. This study presents not only the development of a credit scoring model based on the optimized HistGradientBoosting algorithm but primarily the implementation of a comprehensive automated monitoring architecture deployed on cloud computing services. Using a historical sample of 50,206 records, a workflow integrating feature engineering and cross-validation was designed, achieving a Gini coefficient of 67.4%. The main contribution lies in the deployment of an interactive web platform that enables real-time monitoring of performance and stability metrics. Through this system, the model's consistency was evaluated using the Population Stability Index (PSI), obtaining a value of 0.4%, which confirms the absence of predictive degradation or structural changes in variables over time. This approach demonstrates how the integration of Machine Learning Operations (MLOps) techniques and advanced visualization tools facilitates continuous auditing and early detection of drifts, surpassing the limitations of traditional static reporting.

KEYWORDS: credit risk / machine learning / model monitoring / model risk management

## INTRODUCCIÓN

En la actualidad, los modelos de inteligencia artificial (IA) están presentes en todos los sectores económicos. El sector bancario peruano no es ajeno a esta tendencia. Su importancia ha sido tal que se han emitido normas que promueven el uso de la inteligencia artificial en el sistema financiero del país (Superintendencia de Banca, Seguros y AFP [SBS], 2025). El uso de la IA se refleja en aumentos concretos de productividad. Estudios han documentado que la adopción de IA en empresas bancarias creció un 100 % durante la pandemia (Aguirre Felix Díaz et al., 2022); sin embargo, este crecimiento trajo consigo cuestionamientos sobre sesgo, discriminación, transparencia e interpretación que pueden afectar directamente a los usuarios (Carbó-Valverde & Rodríguez-Fernández, 2024).

Antes de la llegada masiva del *machine learning*, el sector bancario usaba algoritmos estadísticos fáciles de interpretar. El *score* crediticio era producto de una regresión logística, cuyos coeficientes podían analizarse directamente. Los modelos de *scoring* clásicos catalogan a los clientes según su nivel de riesgo en función de un conjunto de datos (Enriquez Montes & Orbezo Huamancari, 2022). Con la llegada del *machine learning*, quedó claro que estos algoritmos superan a los métodos tradicionales: la *support vector machine* (SVM) detectó un 20 % más de clientes riesgosos que la regresión logística (Cervantes Artavia et al., 2023), mientras que *random forest* junto con redes neuronales superaron a la regresión en 17 y 20 puntos de *area under the curve* (AUC), respectivamente (Valenzuela Sánchez, 2025).

Esta realidad llevó a la industria bancaria a adoptar algoritmos como XGBoost, KNN, Naive Bayes y Decision Tree, que forman parte del estado del arte en modelos de *scoring* (Tocto Reyes et al., 2022). Una ventaja central de estos métodos es que capturan relaciones no lineales entre variables, adaptándose mejor a un entorno cambiante (Cosme Cervera et al., 2025). A pesar de ello, esta ganancia en rendimiento tiene un costo: los modelos se vuelven difíciles de interpretar.

Frente a esto, la SBS emitió un reglamento que obliga a las instituciones a garantizar la gestión del riesgo de modelos dentro de las organizaciones bancarias, de seguros y de las AFP. Las debilidades en el desarrollo y seguimiento de modelos, como metodologías inadecuadas o supuestos incorrectos, pueden generar pérdidas y consecuencias adversas (Resolución S.B.S. 00053-2023). Para responder a estas exigencias, surgieron técnicas de explicabilidad como *shapley additive exPlanations* (SHAP) y *local interpretable model-agnostic explanations* (LIME), que permiten identificar el aporte de cada variable dentro de un algoritmo complejo (García García, 2024). Complementariamente, un correcto análisis exploratorio de datos y la ingeniería de variables facilitan la interpretabilidad desde la base del modelo (Erazo Laínez, 2024).

No obstante, entrenar bien un modelo no es suficiente. Una vez desplegado en producción, su comportamiento cambia conforme cambia la población que analiza. Por

lo tanto, el seguimiento debe cubrir dos frentes: el rendimiento, que mide qué tan bien distingue el modelo, y la estabilidad, que detecta variaciones en la predicción y en las variables predictoras (Cordova Benitez et al., 2024). Para evaluar el rendimiento se usan métricas como el coeficiente de Gini y las derivadas de la matriz de confusión (Dos Santos et al., 2025). Para evaluar la estabilidad, el índice de estabilidad poblacional (*population stability index*, PSI) permite medir el cambio en la distribución de la población respecto de una referencia y definir umbrales de alerta ante posibles cambios en la distribución (*drifts*) (Reyes Morales & Sosa Castro, 2022).

A pesar de los avances alcanzados por separado en cada uno de estos frentes, existe una brecha concreta en la literatura: los estudios de *scoring* se enfocan en el desarrollo del modelo, los de explicabilidad lo tratan como un análisis estático y los de *machine learning operations* (MLOps) abordan el monitoreo como un problema de infraestructura genérica. Ninguno integra estos tres componentes dentro de un marco orientado al cumplimiento regulatorio del sistema financiero peruano. Entonces, este artículo propone un flujo de trabajo que cubre el desarrollo, la explicabilidad mediante SHAP y el monitoreo automatizado en producción, considerando los lineamientos de la Resolución 00053-2023 de la SBS. Se comparan múltiples algoritmos con validación cruzada, se aplica un ajuste de hiperparámetros y se despliega una plataforma web de seguimiento en tiempo real que toma el conjunto de entrenamiento como referencia para detectar desviaciones futuras.

## ESTADO DEL ARTE

### Modelos de *machine learning* para riesgo crediticio

La predicción del incumplimiento crediticio ha dejado de depender exclusivamente de la regresión logística. Noriega et al. (2023), en una revisión de cincuenta y dos estudios sobre riesgo crediticio en microfinanzas, documentaron que los modelos de tipo *boosting* (XGBoost, AdaBoost, LightGBM) son los más efectivos y los más investigados, cuyos principales obstáculos recurrentes son el desequilibrio de clases y la baja representatividad de los datos. Por su parte, Suhadolnik et al. (2023) confirmaron esta tendencia con evidencia empírica sólida: evaluaron diez algoritmos sobre una cartera con más de 2,5 millones de registros y encontraron que los modelos de ensamble, en particular XGBoost, superan ampliamente a los métodos tradicionales en la clasificación de créditos.

En línea con lo anterior, Rahmani et al. (2024) propusieron un flujo de trabajo sistemático que incorpora la codificación *weight of evidence* para el tratamiento de valores atípicos y faltantes, junto con la optimización de hiperparámetros mediante algoritmos genéticos multiobjetivo, con lo que se obtuvieron modelos más robustos desde el punto de vista financiero. Yang et al. (2024), por su parte, aplicaron ensamble de modelos con LightGBM y XGBoost sobre el dataset Home Credit, a fin de gestionar el desbalance con SMOTE y mostrar que las variables de historial crediticio externo dominan la predicción.

En el contexto peruano, Castañeda Hilario et al. (2025) desarrollaron un modelo explicable para tarjetas de crédito de bancos como BBVA e Interbank mediante LightGBM e integraron SHAP, con un AUC de 0,94. Con ello, se identificó el uso intensivo de la línea de crédito y la morosidad acumulada como los predictores de mayor peso.

### Explicabilidad e inteligencia artificial responsable

El alto rendimiento predictivo de los modelos complejos se enfrenta a una exigencia regulatoria concreta: las instituciones deben poder explicar sus decisiones. Hermitaño Castro (2022), en una revisión sistemática publicada en esta misma revista, señaló que la caja negra es la desventaja central de los algoritmos de *machine learning* (ML) en crédito y recomendó el uso de SHAP como vía para transparentar los resultados sin sacrificar rendimiento. De Lange et al. (2022) demostraron esta afirmación en un banco noruego real: un modelo LightGBM combinado con SHAP superó al modelo logístico interno del banco, identificando la volatilidad del saldo de crédito y la duración de la relación con el cliente como variables determinantes, y facilitando la interpretación ante reguladores. Weber et al. (2024), en una revisión de sesenta artículos sobre XAI en finanzas, observaron que el campo ha migrado de modelos inherentemente transparentes hacia técnicas de explicabilidad *post-hoc* sobre algoritmos más complejos, de las cuales SHAP fue la más utilizada. Cepeda Cavero y Véliz Soto (2025) analizaron las brechas entre principios éticos y gobernanza algorítmica real en la banca entre 2020 y 2025, y encontraron que reducir el sesgo puede costar hasta un 9 % de precisión y que la equidad estadística no siempre coincide con la equidad percibida por el usuario. Estos hallazgos refuerzan que la explicabilidad no es solo técnica sino también organizacional y regulatoria.

### Monitoreo de modelos y MLOps

Entrenar un modelo no es suficiente, pues su comportamiento en producción cambia conforme cambia la población que analiza; además, detectar esa deriva a tiempo es parte del ciclo de vida del modelo. Dar y Cavus (2026) propusieron el método *profile drift detection* (PDD), que usa perfiles de dependencia parcial para identificar no solo cuándo el modelo se deteriora, sino también qué variable ha cambiado su relación con el objetivo, integrando XAI directamente al monitoreo. Bhagavathula (2025) desarrolló una plataforma de observabilidad en tiempo real basada en OpenTelemetry compatible con AWS, Azure y GCP, que detecta la deriva con una precisión de 94,2 % y menos de 5 % de falsos positivos combinando pruebas estadísticas como KS, PSI y chi-cuadrado de forma simultánea. Shcherbinin (2026) abordó el reentrenamiento adaptativo en entornos industriales mediante la orquestación con Apache Airflow y redujo el tiempo de detección de incidentes en un 60 % y la carga manual de preparación de datos en un 40 %. Estos trabajos muestran que el monitoreo efectivo no se reduce a calcular una métrica periódica, sino que requiere una arquitectura tecnológica diseñada para operar en producción de forma continua y auditable.

### Marco regulatorio

El contexto normativo es inseparable del problema técnico. En el Perú, la Superintendencia de Banca, Seguros y AFP emitió la Resolución 00053-2023, que establece el Reglamento de Gestión de Riesgos de Modelo y obliga a las instituciones financieras a garantizar la estabilidad, la interpretabilidad y el monitoreo de sus modelos en producción (Zegarra Jara, 2025). Internacionalmente, el marco SS1/23 de la Autoridad de Regulación Prudencial del Reino Unido (True North Partners, 2023) clasifica el riesgo de modelo como una disciplina independiente y exige validación independiente, gobernanza liderada por el directorio y controles sobre el desarrollo y la implementación. Joshi (2025) advierte que los marcos MRM tradicionales no son suficientes para los modelos de IA actuales, dado que sus características estocásticas y la falta de trazabilidad dificultan la validación estándar, y propone integrar simulaciones de Monte Carlo y alinearse con estándares como NIST AI RMF e ISO/IEC 23894. Este estudio responde directamente a estos tres niveles regulatorios: local, regional e internacional.

La revisión de la literatura permite identificar una brecha concreta. Los estudios sobre modelos de ML para crédito se concentran en el desarrollo y en la comparación de algoritmos; los trabajos sobre XAI abordan la explicabilidad como un análisis estático posterior al entrenamiento; y los de MLOps tratan el monitoreo como un problema de infraestructura genérica. Ninguno de estos ejes se ha integrado en un solo flujo de trabajo orientado al cumplimiento regulatorio específico del sistema financiero peruano. Este estudio cubre esa brecha al proponer un marco que conecta el desarrollo del modelo, la explicabilidad mediante SHAP y el monitoreo automatizado en producción dentro de los lineamientos de la Resolución 00053-2023 de la SBS, lo que convierte el ciclo de vida del modelo en una unidad auditable y reproducible.

## METODOLOGÍA

En este apartado se explicará a detalle las fases del desarrollo y seguimiento del modelo propuesto.

### *Dataset*

Para el presente estudio se tomó una muestra de clientes del portafolio de una institución financiera peruana interesada en analizar el comportamiento de impago de su cartera. Los periodos disponibles comprenden desde el 202301 hasta el 202403, con un total de 50 206 registros y 44 variables que combinan información externa del reporte crediticio consolidado (RCC) con información interna. El reporte RCC permite conocer la salud financiera de un individuo natural o de una persona jurídica en las diferentes entidades del sistema financiero, lo que otorga una visión más amplia para el análisis (Romero Carmen, 2024). En la Tabla 1, se muestran los grupos de variables disponibles y su descripción general.

**Tabla 1**

*Conjunto de variables del* dataset *de la institución financiera*

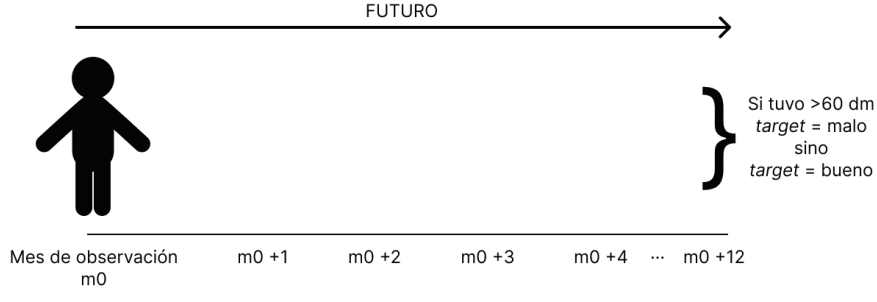
| Fuente | # variables | Descriptivo general de variables |
|---|---|---|
| *Stock* | 5 | Diferencia de meses de buen y mal comportamiento. DIF_BUE_MA |
| RCC | 1 | *Flag* deuda vencida últimos 24 meses. FLG_DVNCD_24M |
| RCC | 3 | Incremento de deuda / incremento de entidades. INC_X |
| *Stock* | 4 | Máximo número de días con atraso. MAX_ATRA_X |
| RCC | 3 | Máxima deuda reportada / máximas entidades. MAX_DEU_ENT_X |
| *Stock* | 4 | Numero de meses con atrasos mayores a X días. NMES_UATR_X |
| *Stock* | 5 | Meses con buen comportamiento. N_BU_X |
| RCC | 7 | Numero de meses con deuda o normal o diferente normal. N_DEU_X |
| RCC | 9 | Máximos, promedios y ratios de deudas activas o hipotecarias |
| RCC | 3 | Variaciones de calificaciones, deudas y entidades. VAR_X |

*Nota*. Las variables fueron agrupadas por relación. Esto no significa que sean exactamente iguales, ya que tienen diferentes ventanas de tiempo y casuísticas especiales.

En cuanto a los aspectos éticos, el *dataset* utilizado fue extraído en el marco de las funciones del autor dentro de la institución financiera y se encuentra completamente anonimizado, con los identificadores personales encriptados, por lo que ningún registro puede vincularse a un individuo específico. El estudio no involucra recolección de datos primarios ni contacto directo con personas, por lo que no requiere aprobación de un comité de ética. Dado que los modelos de *scoring* crediticio pueden afectar el acceso de personas a productos financieros, este trabajo enfatiza la interpretabilidad y explicabilidad del modelo como mecanismos para reducir el riesgo de decisiones arbitrarias o sesgadas, en línea con los principios de gobernanza algorítmica responsable (Cepeda Cavero & Véliz Soto, 2025).

**Figura 1**

*Diagrama conceptual de la construcción del* target




*Nota*. Para la creación del *target* se toma el máximo días de mora en los siguientes doce meses.

## Análisis exploratorio de datos (EDA)

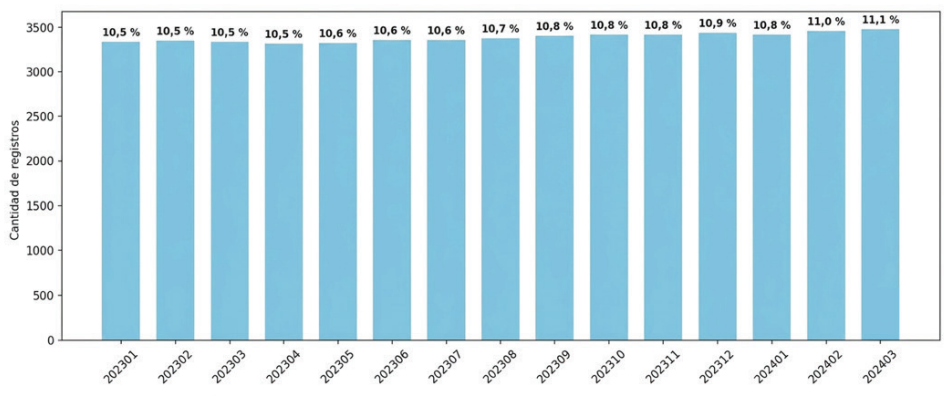
### *Concepto y estabilidad de la variable objetivo (*target*)*

Para este artículo se definió que el *target* tenga dos posibles valores: cliente con *default* (malo) y cliente sin *default* (bueno). Los clientes malos son aquellos que tienen más de sesenta días de mora en los meses siguientes. En la Figura 1, se muestra en detalle el concepto del *target*.

Una vez definido el *target*, es importante analizar su distribución y su estabilidad a lo largo del tiempo. En la Figura 2, se muestra la cantidad de registros y la tasa de malos (TM) para cada uno de los periodos. La tasa de malos se define como el porcentaje de malos respecto del total de clientes; este indicador refleja la calidad de la cartera y puede presentar valores fluctuantes según el tipo de *stock* que se analice (Reyes Morales & Sosa Castro, 2022). Como se puede observar, el *dataset* se considera desbalanceado, con una proporción de 10 a 1. Este problema será abordado más adelante con hiperparámetros internos del modelo final, así que no es necesario aplicar una técnica de *over-/subsampling*.

**Figura 2**

*Evolutivo de tasa de malos y cantidad de clientes*




*Nota*. Se aprecia que la tasa de malos es estable a lo largo de los periodos.

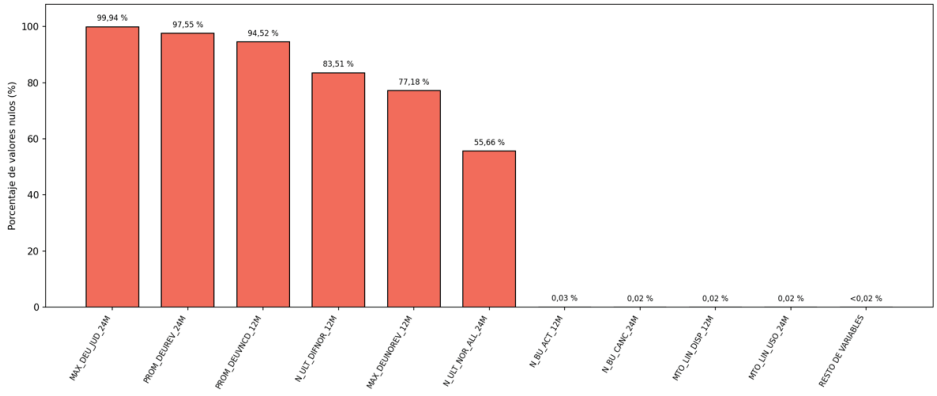
### *Análisis de valores faltantes en las variables*

Durante esa fase, se analizaron las 44 variables, a fin de reconocer cuáles de ellas presentaban un alto nivel de valores faltantes (*missings*). En el caso de que las variables presentaran más del 90 % de *missings*, se retiraban del *dataset*, ya que su aporte será ínfimo al presentar grandes cantidades de valores en blanco. Asimismo, los *missings*

pueden imputarse; sin embargo, al tener una gran concentración de valores nulos, su efecto puede decantar en entrenar un modelo de manera incorrecta. Imputar variables con muchos *missings* puede causar sesgos y relaciones incorrectas (Moneta Pizarro et al., 2022). En la Figura 3, se muestran tres variables que poseen más del 90 % de *missings*, las cuales serán retiradas del *dataset*, por lo que tras este primer filtro, permanecen 41 de las 44 variables iniciales.

**Figura 3**

*Porcentaje de* missings *en las cuarenta y cuatro variables iniciales*




*Nota*. Figura creada con la librería missingno y matplotlib en Python.

### *Poder predictivo de variables*

Para no entrenar el modelo con variables innecesarias, es importante realizar un correcto filtro sobre la base de su importancia. Una de las métricas más conocidas para medir el impacto y la relación que tienen las variables con respecto al *target* es el coeficiente de Gini. Esta métrica es la más usada, ya que su interpretación es sencilla debido a que indica el grado de diferenciación que se establece según las hojas de un árbol. Sus valores oscilan entre -1 y 1; los valores más extremos indican una mejor relación indirecta o directa, respectivamente (Gantman, 2020). En la Figura 4, se observan aquellas variables que no pasaron el filtro de importancia. Para el presente estudio, se eliminaron aquellas variables con un coeficiente de Gini menor del 5 % en valor absoluto.

**Figura 4**

*Variables con coeficiente de Gini menor al 5 %*

| | variable | gini |
|---|---|---|
| **33** | NMES_UATR15_I_U6K_24M | 0.043617 |
| **34** | N_NEG_24M | 0.042934 |
| **35** | INC_MAX_DTOTSH_ACTU_24M | 0.036731 |
| **36** | N_DEU_ACTSH_18M | 0.035886 |
| **37** | INC_MAX_ENT_ACTU_24M | 0.031567 |
| **38** | RMAX_DREV_DDIR_24M | 0.019963 |
| **39** | R_DREV_DDIR_24M | 0.019903 |
| **40** | MAX_DEUNOREV_12M | 0.017121 |

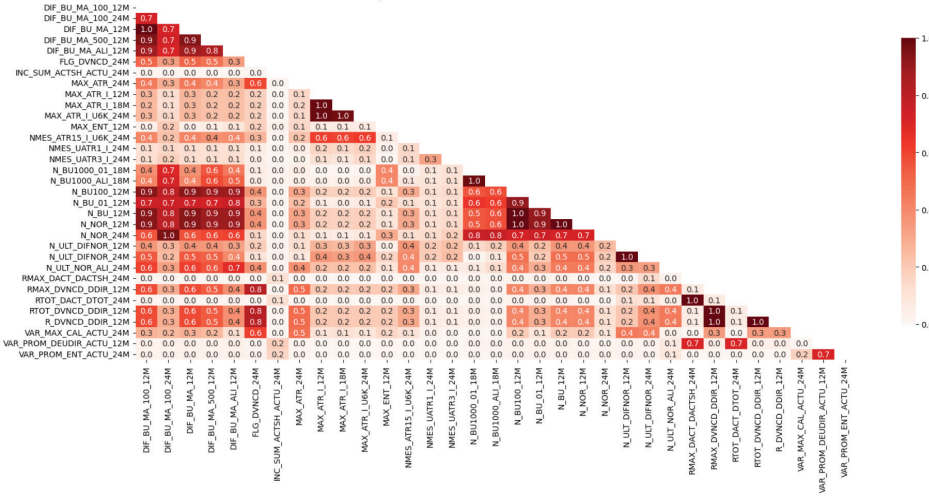
*Nota*. Ocho variables fueron eliminadas del *dataset* debido a su bajo poder predictivo.

### *Correlación de variables*

Hasta este punto se ha analizado cada variable de manera individual, pero se ha obviado una parte fundamental: estas variables pueden estar relacionadas, por lo que se estaría incurriendo en redundancia al incluir variables correlacionadas entre sí. Frente a ello, la correlación de Pearson nos permite conocer la dirección y la intensidad de la relación entre variables (Pérez Monsalve, 2024). Debido a esto, en la Figura 5, se presenta el mapa de calor de las variables restantes en función a la correlación de Pearson.

**Figura 5**

*Mapa de calor con correlaciones para las variables*

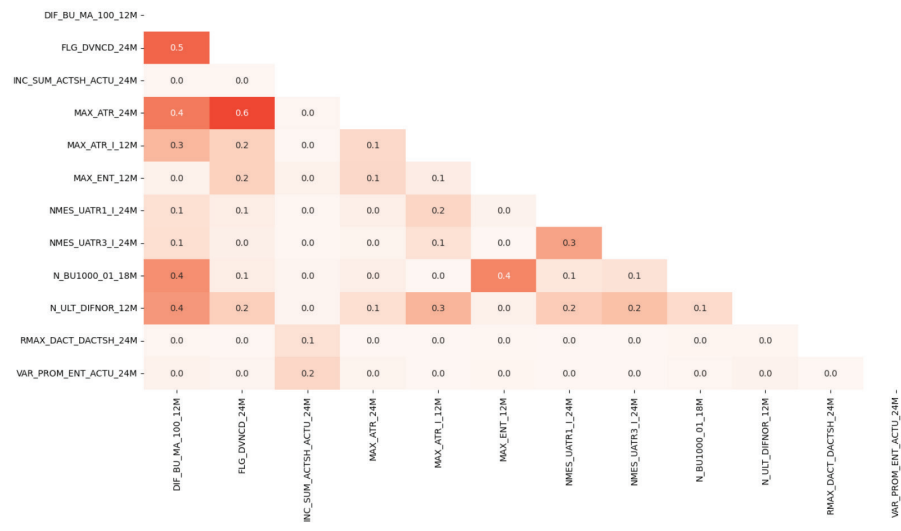



*Nota*. Los valores mostrados son absolutos para una mejor compresión del grado de correlación.

El mapa de calor (Figura 5) muestra que muchas variables están correlacionadas entre sí, por lo que es necesario fijar un umbral para eliminar la redundancia. Una correlación se considera moderada cuando es menor a 0,5 en valor absoluto (Polo Romero, 2024); sin embargo, se eligió 0,6 por dos motivos. El primero es que bajar a 0,5 hubiera eliminado variables que ya pasaron el filtro del coeficiente de Gini; es decir, variables con poder predictivo comprobado. El segundo es que tener variables muy correlacionadas complica el análisis bivariado, porque, al mostrar tasas de malos similares, se vuelve difícil distinguir qué aporta cada una al negocio. Por estas razones, el umbral de 0,6 equilibra la reducción de redundancia sin perder información relevante, lo que resulta en doce variables finalistas (Figura 6).

**Figura 6**

*Mapa de calor con correlaciones para las variables finalistas*




*Nota*. Estas variables representan el 27 % del set inicial luego de pasar por los diversos filtros establecidos.

## Ingeniería de variables

En esta etapa se realizó un último análisis para observar la relación bivariada que existe entre las variables finalistas y el *target*; de ser el caso, sobre la base de este análisis, se realizó una imputación de datos con base en criterios de experto para las variables que contienen *missings*.

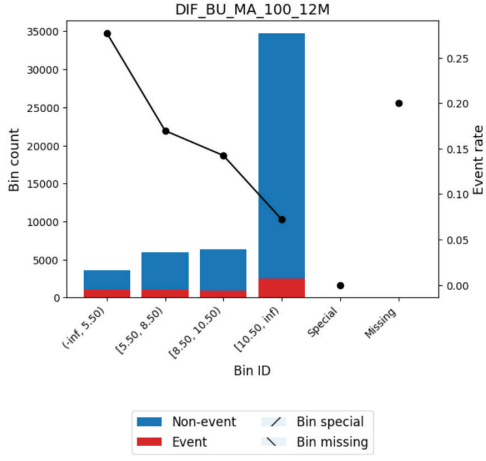
### *Análisis bivariado*

Uno de los problemas que tiene el uso de algoritmos de *machine learning* es entender cómo interactúan las variables con la probabilidad predicha. Debido a esto, es fundamental

realizar un análisis para comprender el sentido que tienen; es decir, si premian o castigan la probabilidad de impago. El *binning* es una de las técnicas más usadas para agrupar variables, además nos permite conocer, de manera sencilla, el sentido monótono que tienen en el riesgo crediticio (Wheeler & Carrol, 2023). En las figuras 7 y 8, se pueden ver los cortes óptimos tomados por la librería a través del uso de un algoritmo de programación convexa que maximiza el *information value* (IV), bajo la hipótesis de que la variable es monótona.

**Figura 7**

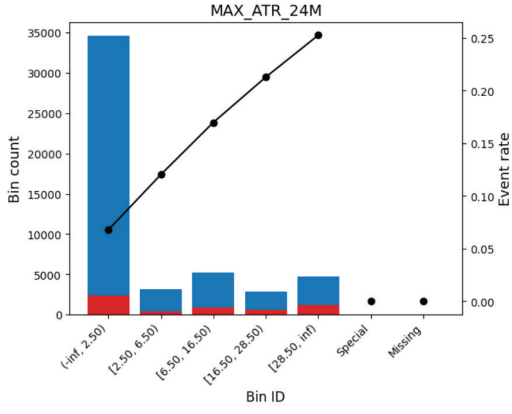
*Gráfico bivariado para la variable DIF_BU_MA_100_12M*




*Nota*. Se observa que, en esta variable relación inversa, a mayor valor de la variable, menor es el riesgo.

**Figura 8**

*Gráfico bivariado para la variable MAX_ATR_24M*




*Nota*. Se observa que, en esta variable relación directa, a mayor valor de la variable, mayor es el riesgo.

De este análisis, se obtuvieron las conclusiones que se observan en la Tabla 2. En esta se muestra la variable, su descripción y el sentido en relación con el *target*.

**Tabla 2**

*Diccionario de variables finalistas*

| Fuente | Descripción | Interpretación | Sentido |
|---|---|---|---|
| DIF_BU_MA_100_12M | Diferencia entre meses de buen y mal comportamiento en los últimos doce meses | A mayor diferencia positiva (más meses buenos), menor riesgo. | Inverso |
| FLG_DVNCD_24M | Flag deuda vencida de los últimos veinticuatro meses | El valor de 1 indica mayor riesgo | Directo |
| INC_SUM_ACTSH_ACTU_24M | Incremento de la deuda total en los últimos veinticuatro meses | Si la deuda crece mucho, el riesgo aumenta; por tanto, el incremento alto empeora el perfil. | Inverso |
| MAX_ATR_24M | Máximo número de días de atraso en los últimos veinticuatro meses | Más días de atraso, mayor riesgo. | Directo |
| MAX_ATR_I_12M | Máximo atraso en los últimos doce meses | Un atraso más severo en el último año incrementa riesgo. | Directo |
| MAX_ENT_12M | Máximo de entidades acreedoras en los últimos doce meses | Más entidades implican mayor apalancamiento y probabilidad de sobreendeudamiento. | Directo |
| NMES_UATR1_I_24M | Meses desde el último atraso ≥ 1 día (últimos veinticuatro meses) | Cuanto más tiempo ha pasado desde un atraso, menor es el riesgo. | Inverso |
| NMES_UATR3_I_24M | Meses desde el último atraso ≥ 3 días (últimos veinticuatro meses) | Más meses sin atrasos relevantes, menor riesgo. | Inverso |
| N_BU1000_01_18M | Meses de buen comportamiento (últimos dieciocho meses) | Más meses sin atrasos reflejan mejor disciplina de pago y menor riesgo. | Inverso |
| N_ULT_DIFNOR_12M | Meses desde la última clasificación distinta a "normal" (últimos doce meses) | Más tiempo sin una mala clasificación indica recuperación y menor riesgo. | Inverso |
| RMAX_DACT_DACTSH_24M | Ratio entre la máxima deuda activa y la máxima deuda activa sin hipoteca (últimos veinticuatro meses) | Una mayor proporción implica niveles más altos de deuda activa, mayor riesgo | Directo |
| VAR_PROM_ENT_ACTU_24M | Variación en el número de entidades acreedoras (últimos dieciocho meses) | Una reducción en entidades indica des apalancamiento, aumento implica mayor riesgo. | Inverso |

*Nota*. El orden directo indica que, a mayor valor, mayor es el riesgo, mientras que el inverso indica que, a mayor valor, menor es el riesgo, y viceversa.

### *Imputación de variables*

Durante esa sección se detectó que existía un porcentaje ínfimo de variables con *missings* dentro de las variables finalistas (<1 %). Debido a esto se realizó una imputación por

criterio de experto, ya que los *missings* tienen un significado a nivel de negocio. En la Tabla 3, se aprecian las imputaciones realizadas sobre nuestras variables finalistas.

**Tabla 3**

*Imputaciones de variable por criterio de experto*

| Grupo de variables | Fundamento | Valor imputado |
|---|---|---|
| Variables de comportamiento (atrasos, entradas, variaciones) | Sin ocurrencia de evento, sin registros. Ausencia de comportamiento riesgoso | Se imputa por el valor de 0, mantiene lógica de riesgo bajo. |
| Variables de tiempo o frecuencia | Para clientes nuevos | Imputado por 99, valor alto. Sin eventos. |

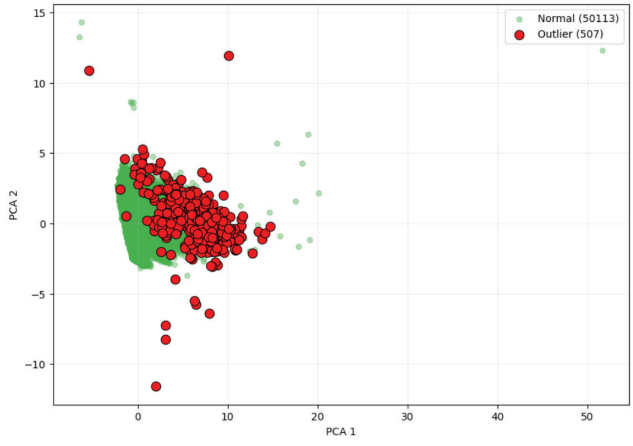
*Nota*. Tras realizar esta imputación tenemos un *dataset* con *missings* de 0 % para todos los valores.

### *Detección de* outliers

Otro paso importante para contar con un *dataset* correcto es identificar los valores *outliers*. Los valores atípicos pueden tener un efecto significativo a la hora de obtener resultados (Dash et al., 2023). Para identificar estos valores atípicos, existe una multitud de técnicas apoyadas en conceptos estadísticos y otras más avanzadas que usan modelos de inteligencia artificial. Dentro de las técnicas más usadas y confiables, se cuenta con la de Isolation Forest. Este algoritmo no supervisado permite aislar y detectar *outliers*, y se caracteriza por su *performance* y confiabilidad en grandes conjuntos de datos (Banchero et al., 2021). En la Figura 9, se puede observar el número de *outliers* detectados en todo nuestro conjunto de entrenamiento, teniendo en cuenta que el grado de contaminación indicado es del 1 %.

**Figura 9**

Outliers *detectados en* dataset

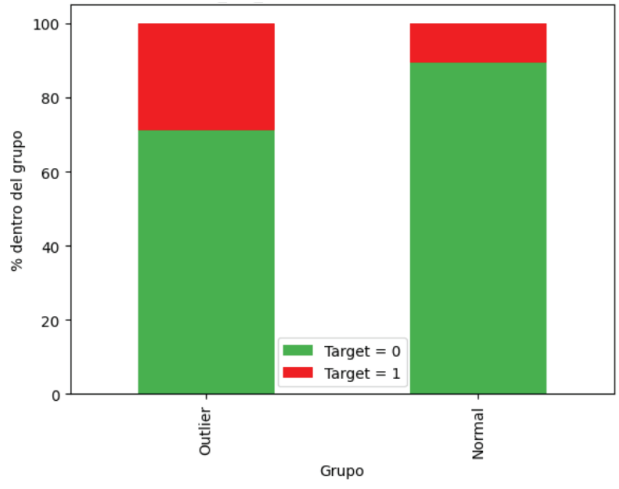



*Nota*. El gráfico mostrado se elaboró a partir de un aplanamiento de análisis de componentes principales (PCA) para visualizar en 2D los valores atípicos.

Si bien los algoritmos basados en árboles de decisión presentan una tolerancia natural a los valores atípicos, debido a que sus particiones se basan en ordenamiento y no en distancias, la eliminación de *outliers* responde a un criterio de calidad del *dataset*. En datos financieros, los valores atípicos suelen corresponder a errores de registro o comportamientos que no representan a la población objetivo de la cartera analizada, por lo que su inclusión puede introducir ruido en las relaciones entre variables y distorsionar el análisis bivariado. La Figura 10 respalda esta decisión, pues muestra que la distribución de malos y buenos es similar entre el grupo de *outliers* y el grupo normal, lo que confirma que su retiro no genera un sesgo en la variable objetivo y que el *dataset* resultante mantiene la representatividad de la cartera.

**Figura 10**

*Distribución de* target *en* outliers




*Nota*. Las distribuciones de malos rondan el 30 % y el 10 % para grupos *outliers* y normales, respectivamente. No existen grandes diferencias en su distribución.

### Entrenamiento de modelo

Para establecer el conjunto de *train*/validación, se consideraron los meses comprendidos desde el 202301 hasta el 202312, mientras que para la base de fuera tiempo (*out of time*) se utilizó la data correspondiente del periodo 202401-202403. Con el fin de obtener resultados transparentes, se aplicó validación cruzada estratificada con un pliegue de *folds* igual a 5. La validación cruzada permite mejorar la robustez del modelo, lo que favorece su aplicación (Zuleta Fuerte et al., 2025). En la Tabla 4, se observan todos los algoritmos entrenados con sus parámetros por defecto, la validación cruzada y las métricas obtenidas.

**Tabla 4**

*Métricas obtenidas para todos los modelos*

| Modelo | AUC_CV_ mean (%) | Gini_CV_ mean (%) | KS_CV_ mean (%) | AUC_ OOT (%) | Gini_ OOT (%) | KS_OOT (%) | Variación Gini |
|---|---|---|---|---|---|---|---|
| HistGradient-Boosting | 82,9 | 65,8 | 51,0 | 83,7 | 67,3 | 52,0 | 0,1791 |
| LightGBM | 82,3 | 64,7 | 50,6 | 83,0 | 66,0 | 51,1 | 0,1829 |
| GradientBoost-ing | 82,5 | 65,0 | 50,5 | 82,8 | 65,6 | 50,3 | 0,1778 |
| CatBoost | 82,0 | 64,0 | 50,4 | 82,4 | 64,8 | 50,9 | 0,1844 |
| AdaBoost | 81,9 | 63,8 | 49,8 | 82,4 | 64,7 | 49,9 | 0,1862 |
| LogisticRegres-sion | 81,4 | 62,7 | 48,8 | 81,7 | 63,5 | 48,4 | 0,1900 |
| XGBoost | 80,5 | 61,0 | 47,6 | 81,4 | 62,8 | 48,7 | 0,2045 |
| SVC | 81,1 | 62,2 | 50,4 | 81,4 | 62,8 | 49,7 | 0,1923 |
| SGDClassi-fier_log | 78,0 | 56,1 | 43,6 | 80,6 | 61,1 | 45,9 | 0,2447 |
| RandomForest | 79,8 | 59,6 | 46,9 | 80,0 | 60,0 | 46,6 | 0,2043 |
| ExtraTrees | 78,7 | 57,5 | 45,4 | 79,3 | 58,6 | 45,1 | 0,2183 |
| BernoulliNB | 77,0 | 54,1 | 45,4 | 77,3 | 54,7 | 44,2 | 0,2329 |
| GaussianNB | 76,3 | 52,7 | 42,6 | 76,9 | 53,9 | 43,7 | 0,2424 |
| Bagging_DT | 75,4 | 50,7 | 41,8 | 75,2 | 50,4 | 45,0 | 0,2448 |
| KNN | 72,5 | 44,9 | 40,0 | 71,4 | 42,7 | 45,8 | 0,2645 |
| DecisionTree | 61,3 | 22,6 | 23,7 | 59,5 | 18,9 | 43,9 | 0,3684 |

*Nota*. Las métricas CV hacen referencia a las *cross validation,* mientras que la muestra OOT es el periodo que nunca se tomó en cuenta a la hora de entrenar/validar.

Según las métricas obtenidas, se observó que el algoritmo HistGradientBoosting obtuvo los mejores indicadores: un mejor rendimiento de casi el 2 % por encima del coeficiente de Gini en su OOT respecto del conjunto CV. Este algoritmo potencia su forma original de *boosting* utilizando la técnica del *binning*, lo que reduce la complejidad computacional y mejora la eficiencia del modelo (Ferro Rugeles et al., 2023). Además, este algoritmo se caracteriza por presentar una gran robustez y controlar de manera eficiente las interacciones entre variables. Como lo señala Betancourt Ludeña (2025), este algoritmo puede superar inclusive a redes neuronales profundas en problemas de clasificación y gana no solo en *performance*, sino también en interpretabilidad. En el presente estudio se optó por algoritmos de *machine learning* y no de *deep learning*, precisamente porque se quiso mejorar el concepto de interpretabilidad a la hora de implementar el modelo.

### Optimización de hiperparámetros

Un paso crítico es ajustar (optimizar) el modelo. En la sección anterior comparamos los modelos bajo el supuesto de los hiperparámetros por defecto. Optuna es una librería que realiza una búsqueda inteligente para encontrar la mejor configuración de hiperparámetros y se basa en algoritmos de optimización de árboles en búsqueda de bandas (Martínez Martínez, 2024). En la Figura 11, se aprecian los hiperparámetros que se buscaron dentro de la optimización, así como el espacio de búsqueda.

**Figura 11**

*Búsqueda de hiperparámetros mediante Optuna*

```
def objective(trial):
    params = {
        "learning_rate": trial.suggest_float("learning_rate", 0.01, 0.3, log=True),
        "max_iter": trial.suggest_int("max_iter", 100, 500),
        "max_depth": trial.suggest_int("max_depth", 2, 8),
        "min_samples_leaf": trial.suggest_int("min_samples_leaf", 20, 200),
        "max_bins": trial.suggest_int("max_bins", 64, 255),
        "l2_regularization": trial.suggest_float("l2_regularization", 1e-4, 10.0, log=True),
        "random_state": SEED,
        "early_stopping": False,
        "loss": "log_loss"
    }
    kf = StratifiedKFold(
        n_splits=N_SPLITS,
        shuffle=True,
        random_state=SEED
    )
    gini_folds = []
    for train_idx, val_idx in kf.split(X_train, y_train):
        X_tr = X_train.iloc[train_idx]
        X_val = X_train.iloc[val_idx]
        y_tr = y_train.iloc[train_idx]
        y_val = y_train.iloc[val_idx]
        sw_tr = np.where(y_tr == 1, w_pos, w_neg)
        model = HistGradientBoostingClassifier(**params)
        model.fit(X_tr, y_tr, sample_weight=sw_tr)
        proba_val = model.predict_proba(X_val)[:, 1]
        gini = gini_score(y_val, proba_val)
        gini_folds.append(gini)
    mean_gini = float(np.mean(gini_folds))
    return mean_gini
```

*Nota*. El espacio de búsqueda fue definido a partir de criterio experto.

Tras realizar la búsqueda con Optuna en cincuenta escenarios, se obtuvieron los hiperparámetros definidos en la Tabla 5. Si bien cincuenta iteraciones parecen un número reducido, Optuna no realiza una búsqueda aleatoria, sino una optimización bayesiana que aprende de cada evaluación anterior para dirigir la búsqueda hacia zonas de mayor rendimiento del espacio de hiperparámetros (Martínez Martínez, 2024). Esto implica que cincuenta iteraciones son suficientes para alcanzar la convergencia en espacios de búsqueda de dimensión moderada, como el definido en este estudio. Los resultados respaldan esta decisión: la métrica del coeficiente de Gini en la CV mejoró casi un punto porcentual respecto del modelo con parámetros por defecto, mientras que en la muestra OOT se mantuvo en 67,3 %, lo que indica que la búsqueda encontró una configuración estable y de buen rendimiento.

**Tabla 5**

*Mejores hiperparámetros obtenidos por Optuna*

| Parámetro | Descripción | Valor |
|---|---|---|
| learning_rate | Controla la velocidad con la que el modelo aprende en cada iteración; valores más bajos hacen el aprendizaje más gradual. | 0,106 |
| max_iter | Número máximo de iteraciones que realizará el algoritmo para ajustar el modelo. | 125 |
| max_depth | Profundidad máxima de cada árbol; limita la complejidad y ayuda a evitar sobreajuste. | 2 |
| min_samples_leaf | Mínimo de muestras requeridas en una hoja para asegurar que los cortes no se basen en grupos pequeños. | 180 |
| max_bins | Máxima cantidad de bins usados para discretizar valores continuos antes del entrenamiento. | 207 |
| l2_regularization | Penalización L2 que ayuda a estabilizar el modelo y reducir sobreajuste. | 0,.0005228 |

*Nota*. Los hiperparámetros tuneados son los que mayormente se personalizan y son los más relevantes.

Adicionalmente, para el tratamiento del desbalance de clases identificado en el apartado "Dataset" de la sección "Metodología", se configuró el parámetro class_weight = 'balanced' del algoritmo HistGradientBoosting. Este parámetro ajusta automáticamente los pesos de cada clase según su frecuencia en el *dataset*, lo que penaliza más los errores sobre la clase minoritaria (malos) sin necesidad de aplicar técnicas externas de remuestreo como SMOTE u *oversampling*.

### Importancia de variables

También es importante identificar las variables que más contribuyen a la predicción. En tal sentido, el modelo se caracteriza por ser de alto rendimiento, pero de poca interpretabilidad. Una de las técnicas para lograr la explicabilidad de nuestros modelos es SHAP. Esta está basada en la teoría de juegos de manera cooperativa y otorga el valor marginal que tiene cada variable a la hora de predecir (Buergo Lagunas, 2025). Este método tiene un fundamento matemático, por lo que es de implementación rápida y, sobre todo, es muy interpretable. Al usar la técnica de SHAP, la institución financiera que interpreta el modelo es capaz de conocer la influencia de cada variable a la hora de dar un seguimiento (Marchant Contreras, 2022). En la Tabla 6, se muestra el *ranking* de las variables según la importancia SHAP.

**Tabla 6**

*Importancia SHAP del modelo ajustado*

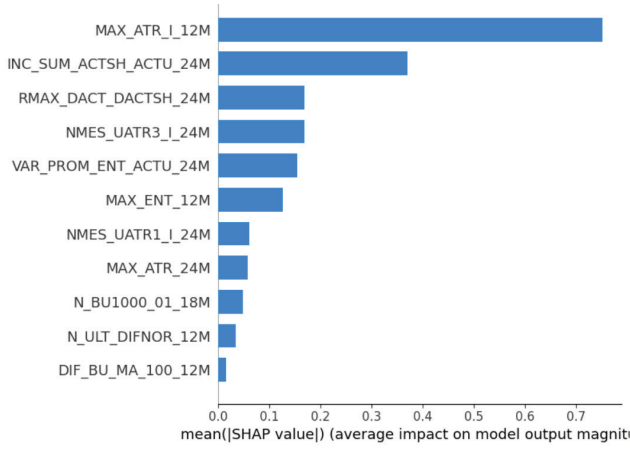
| Variable | Importancia SHAP | SHAP relativo (%) |
|---|---|---|
| MAX_ATR_I_12M | 0,752125 | 38,5 |
| INC_SUM_ACTSH_ACTU_24M | 0,369670 | 18,9 |
| RMAX_DACT_DACTSH_24M | 0,168870 | 8,6 |
| NMES_UATR3_I_24M | 0,168777 | 8,6 |
| VAR_PROM_ENT_ACTU_24M | 0,154431 | 7,9 |
| MAX_ENT_12M | 0,125298 | 6,4 |
| NMES_UATR1_I_24M | 0,059864 | 3,1 |
| MAX_ATR_24M | 0,056815 | 2,9 |
| N_BU1000_01_18M | 0,048253 | 2,5 |
| N_ULT_DIFNOR_12M | 0,033692 | 1,7 |
| DIF_BU_MA_100_12M | 0,014554 | 0,7 |
| FLG_DVNCD_24M | 0 | 0,0 |

*Nota*. El valor de SHAP relativo es el cálculo producto de la suma del valor SHAP entre la suma de todos valores SHAP.

En la Tabla 6, se observa que la variable flag_dvncd_24m no tiene importancia en la predicción, por lo que resulta innecesaria dentro del *dataset*. Debido a esto, se procedió a retirar esta variable y a reentrenar el modelo con los mismos hiperparámetros, pero solo con once variables, conservando las mismas métricas mostradas anteriormente. En la Figura 12, se aprecia el *ranking* de las variables SHAP, acompañado de sus respectivos valores.

**Figura 12**

*Diagrama de barras de importancia SHAP*

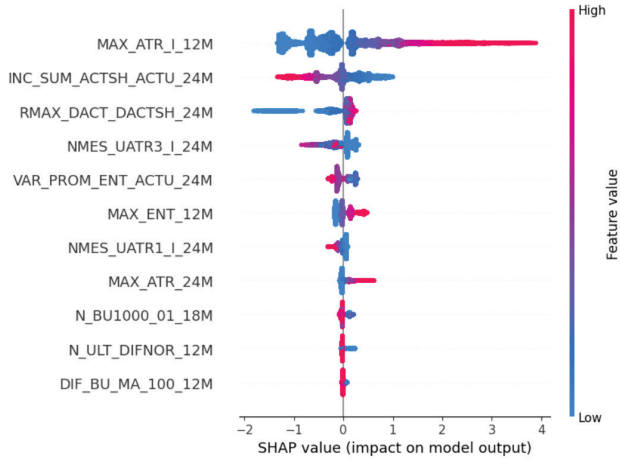



*Nota*. La variable más importante para nuestro modelo es la de máximo días de atraso en los últimos doce meses (top 1).

En la Figura 13, se observa la influencia que tiene cada característica en la predicción de manera resumida, mientras que, en las figuras 14 y 15, se aprecia su contribución con mayor detalle. Los gráficos de dependencia parcial (DP) permiten visualizar el impacto de las variables en la probabilidad, ya que esta técnica usa la contribución individual de cada variable y su relación con la variable objetivo (Jaimes et al., 2023).

**Figura 13**

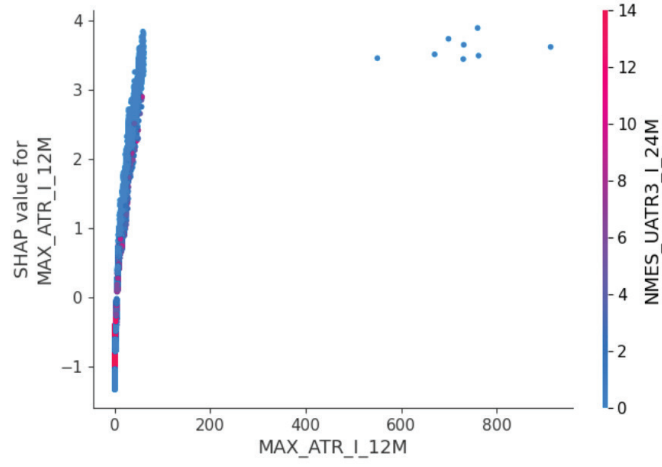
*Diagrama de barras de importancia SHAP*




*Nota*. El color rojo representa mayor probabilidad de riesgo y el azul representa menor probabilidad.

**Figura 14**

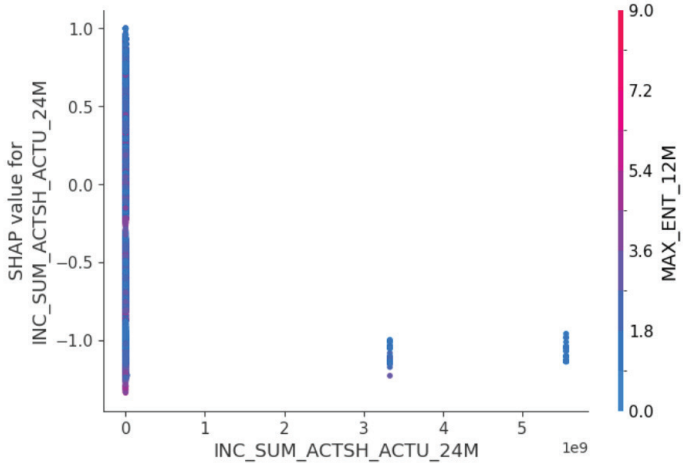
*Diagrama de dependencia parcial variable Max_atr_i_12m*




*Nota*. La figura nos indica que, a mayor valor de la variable, la probabilidad de caer en *default* aumenta.

**Figura 15**

*Diagrama de dependencia parcial variable Inc_sum_actsh_actu_24m*



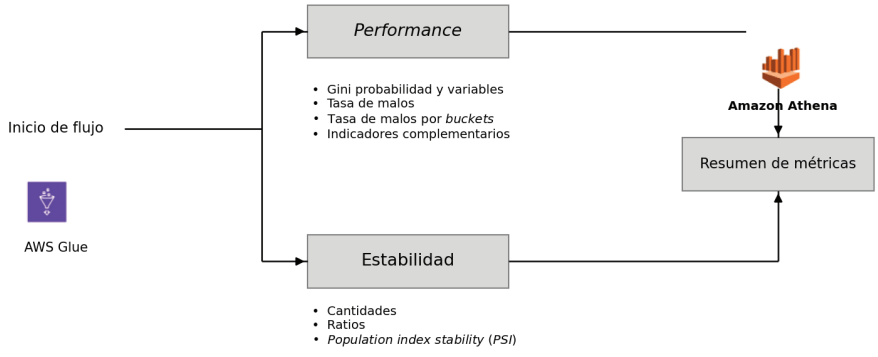

*Nota*. La figura nos indica que, a menor valor de la variable, aumenta la probabilidad de caer en *default*.

## Seguimiento del modelo (monitoreo)

Para poder medir la salud de nuestro modelo, se construyó una serie de *pipelines* que evalúan el modelo en función de dos frentes: el *performance* y la estabilidad de la variable a predecir (probabilidad de *default*) y las variables predictoras. En la Figura 16, se aprecia el flujo seguido para el cálculo de métricas que fue desplegado en ambiente de Amazon Web Services (AWS).

**Figura 16**

*Arquitectura general del flujo para monitoreo*




*Nota*. Los servicios fueron orquestados mediante Step Function.

### Cálculo de indicadores de referencias

Un punto clave para realizar un correcto monitoreo es establecer una referencia, es decir, identificar el *dataset* con el que se estará comparando constantemente el modelo. Para el presente estudio, se escogió como *dataset* de referencia todo el conjunto de *train* del periodo 202301-202312. A partir de este, se tomaron los indicadores de *performance* como el valor base, así como la cantidad y los porcentajes de las variables, los que representan los parámetros utilizados para el seguimiento mensual. En la Tabla 7, se observa el detalle del cálculo de referencia obtenido para realizar el seguimiento/monitoreo del modelo.

**Tabla 7**

*Cálculo de referencia*

| Tipo de monitoreo | Descripción | Valor |
|---|---|---|
| *Performance* | Gini referencia | 68,4 % |
| | Tasa de malos referencia | 10,5 % |
| | Indicadores complementarios | *Acuraccy*: 79<br>Sensibilidad: 73<br>Especificidad: 80 % |
| Estabilidad | *Score*<br>(probabilidad transformada) | Se hace cortes en deciles |
| | Variables | Se hacen cortes con base en los cortes óptimos del *optimal binning* o en su defecto cortes por quintiles |

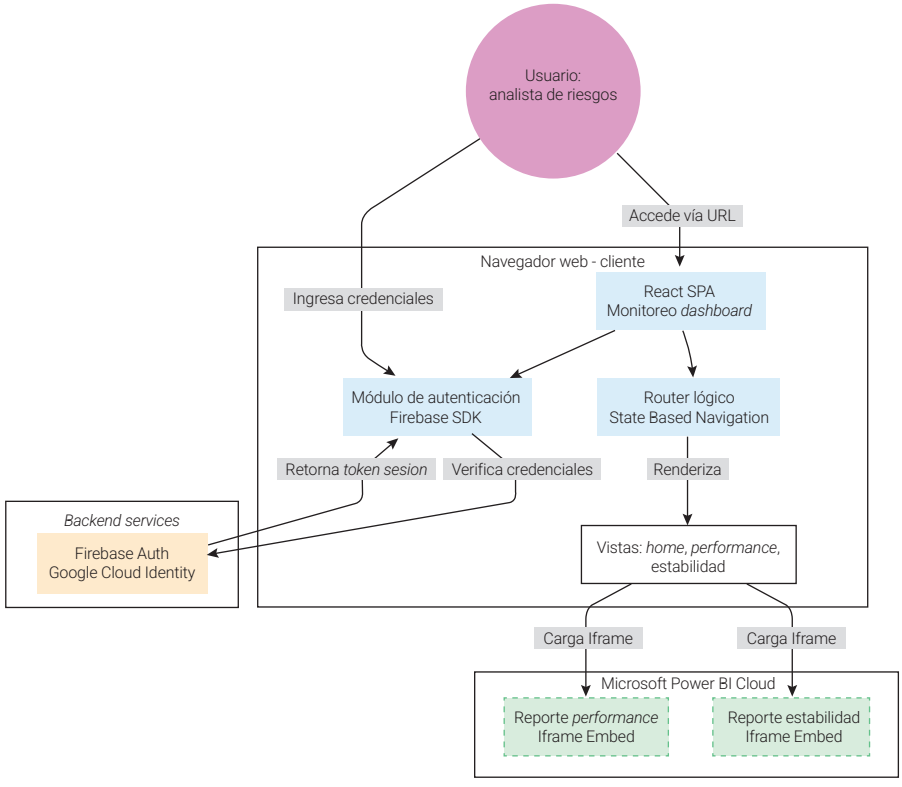
*Nota*. El *score* es la transformación de la probabilidad, representa solo un escalado de su complemento, que oscila entre 0-1000, donde un número mayor indica menor riesgo. Se realizó este escalado para comprender mejor su interpretación.

### Despliegue de la plataforma de seguimiento/monitoreo

Para poder realizar un monitoreo integral del modelo, se desplegó una plataforma que permitiera realizar una interacción directa con los resultados en tiempo real. En la Figura 17, se aprecia la arquitectura de la app web desplegada.

**Figura 17**

*Arquitectura general de la app web desplegada*




*Nota*. La aplicación fue desplegada en Vercel.

## RESULTADOS

Cabe precisar que el presente estudio se enfoca en la evaluación del desempeño estadístico y la estabilidad del modelo, por lo que la cuantificación del impacto financiero, como la estimación de pérdida esperada o beneficios operativos derivados de su implementación, queda fuera del alcance de esta investigación y se plantea como una línea de trabajo futuro.

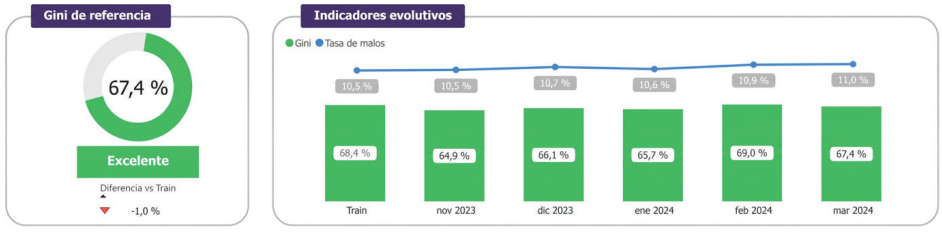
### *Performance* general

Para verificar los resultados obtenidos del modelo, se analizaron los reportes desplegados en la plataforma web de monitoreo. En la Figura 18, se puede observar que, para los últimos meses del *dataset*, el coeficiente de Gini se mantuvo en un 67,4 %. Si comparamos este indicador versus nuestra referencia de *train*, solo es un punto porcentual por

debajo. Este Gini puede ser catalogado como excelente, ya que supera el umbral del 60 % (Putri et al., 2021). Además de tener este indicador sobresaliente, se evidencia que es constante a lo largo del tiempo con ligeras oscilaciones, lo cual indica que es un modelo robusto.

**Figura 18**

*Métricas de* performance *del modelo final*




*Nota*. Diagrama extraído de la plataforma web.

### *Performance* complementario

También es importante verificar que el modelo cuente con una estabilidad en las métricas derivadas de la matriz de confusión (métricas auxiliares). En la Figura 19, se aprecia que las métricas auxiliares se mantuvieron muy similares a lo evidenciado en el *train*, lo que refuerza lo mencionado anteriormente: el modelo cuenta con una robustez importante a lo largo del tiempo. La métrica de *accuracy* que nos indica el porcentaje de acierto del modelo se mantiene en aproximadamente el 80 % en el último mes, con tendencia a subir.

**Figura 19**

*Métricos de auxiliares del modelo final*

**Indicadores de la matriz de confusión**

| Etiqueta_Indicador | Train | nov 2023 | dic 2023 | ene 2024 | feb 2024 | mar 2024 |
|---|---|---|---|---|---|---|
| Accuracy | 79% | 78% | 79% | 79% | 79% | 80% |
| Especificidad | 80% | 79% | 80% | 80% | 80% | 81% |
| F1 Score | 42% | 40% | 41% | 42% | 44% | 43% |
| Odds | 10,5 | 8,4 | 8,8 | 9,9 | 11,1 | 9,9 |
| Prec. Neg. | 96% | 96% | 96% | 96% | 96% | 96% |
| Prec. Pos. | 29% | 28% | 29% | 30% | 31% | 31% |
| Sensibilidad | 73% | 69% | 69% | 71% | 74% | 70% |

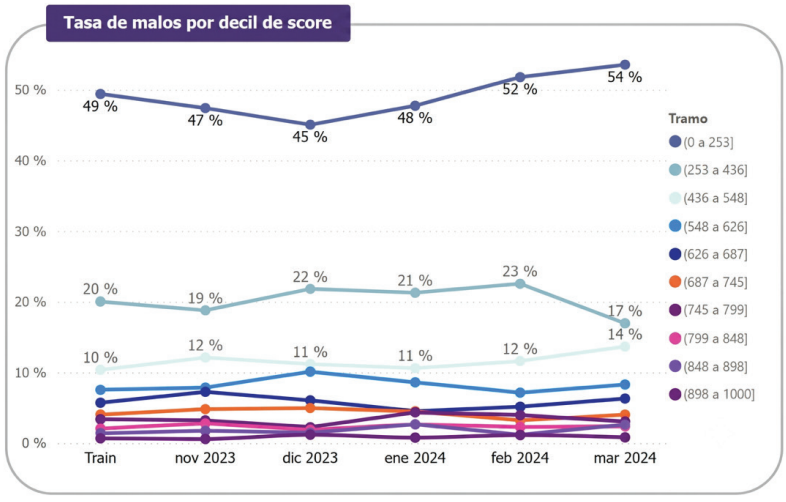
*Nota*. Diagrama extraído de la plataforma web.

### Métricas de negocio

En las secciones anteriores se analizaron métricas de un carácter más estadístico y matemático; sin embargo, no se debe dejar de lado las métricas de negocio. En su mayoría están asociadas a ver si la diferenciación de los clientes predichos como malos y buenos se sigue manteniendo como estaba en *train*, y si, además de esto, se sigue manteniendo un ordenamiento coherente en función del sentido de la variable de *score*. En la Figura 20, se aprecia el evolutivo de la tasa de malos por decil de *score*, ordenados de menor a mayor.

**Figura 20**

*Tasa de malos por decil de cortes de* score

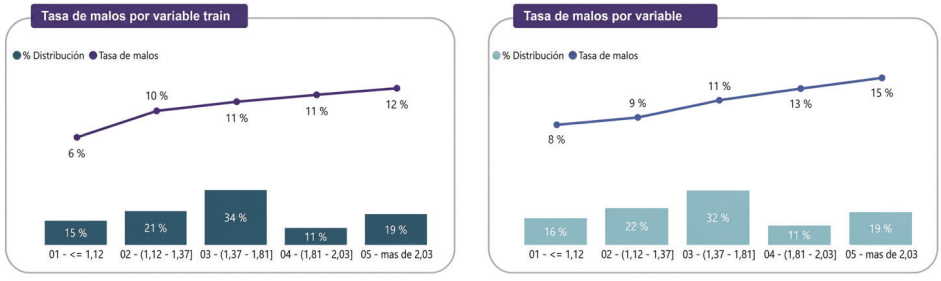



*Nota*. Diagrama extraído de la plataforma web.

Así como se vio el resultado de la predicción (*score*), también es importante analizar cada variable. En la Figura 21, se aprecia la tasa de malos para la variable más importante (top 1) en su periodo de *train* y para el periodo 202403 que corresponde al último monitoreo. Se observa que, a pesar de que han pasado varios meses, la tendencia de la variable mantiene sentido directo, es decir, a mayor valor, mayor probabilidad de riesgo, lo cual es resultado de la correcta selección de variables realizada en las secciones anteriores.

**Figura 21**

*Tasa de malos de la variable top 1 - ratio de deuda máxima y máxima deuda activa sin hipotecario*



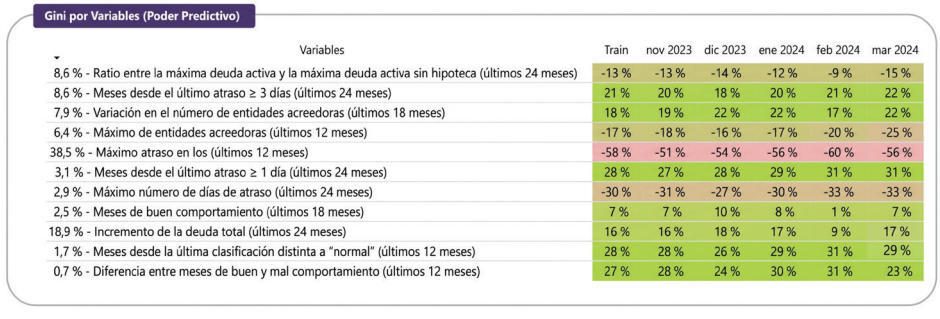

*Nota*. Diagrama extraído de la plataforma web.

## Poder predictivo de variables

Como último frente dentro de la *performance*, también se evaluó que se mantuviera el poder predictivo de las variables. En la Figura 22, se observa que, a lo largo del tiempo, las variables mantuvieron, en su mayoría, valores cercanos a los de *train*, oscilando en hasta ±4 puntos versus su versión de *train*. Esto es un indicador de que el modelo viene performando bien y que no existe ningún cambio de relación con el *target*, lo que garantiza un modelo robusto, interpretable y que refleja confianza en los usuarios finales.

**Figura 22**

*Evolutivo de Gini por variables ordenadas por peso*




| Variables | Train | nov 2023 | dic 2023 | ene 2024 | feb 2024 | mar 2024 |
|---|---|---|---|---|---|---|
| 8,6 % - Ratio entre la máxima deuda activa y la máxima deuda activa sin hipoteca (últimos 24 meses) | -13 % | -13 % | -14 % | -12 % | -9 % | -15 % |
| 8,6 % - Meses desde el último atraso ≥ 3 días (últimos 24 meses) | 21 % | 20 % | 18 % | 20 % | 21 % | 22 % |
| 7,9 % - Variación en el número de entidades acreedoras (últimos 18 meses) | 18 % | 19 % | 22 % | 22 % | 17 % | 22 % |
| 6,4 % - Máximo de entidades acreedoras (últimos 12 meses) | -17 % | -18 % | -16 % | -17 % | -20 % | -25 % |
| 38,5 % - Máximo atraso en los (últimos 12 meses) | -58 % | -51 % | -54 % | -56 % | -60 % | -56 % |
| 3,1 % - Meses desde el último atraso ≥ 1 día (últimos 24 meses) | 28 % | 27 % | 28 % | 29 % | 31 % | 31 % |
| 2,9 % - Máximo número de días de atraso (últimos 24 meses) | -30 % | -31 % | -27 % | -30 % | -33 % | -33 % |
| 2,5 % - Meses de buen comportamiento (últimos 18 meses) | 7 % | 7 % | 10 % | 8 % | 1 % | 7 % |
| 18,9 % - Incremento de la deuda total (últimos 24 meses) | 16 % | 16 % | 18 % | 17 % | 9 % | 17 % |
| 1,7 % - Meses desde la última clasificación distinta a "normal" (últimos 12 meses) | 28 % | 28 % | 26 % | 29 % | 31 % | 29 % |
| 0,7 % - Diferencia entre meses de buen y mal comportamiento (últimos 12 meses) | 27 % | 28 % | 24 % | 30 % | 31 % | 23 % |

*Nota*. Diagrama extraído de la plataforma web.

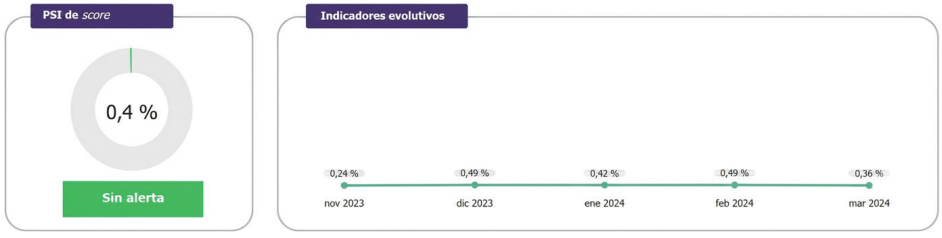
## Estabilidad general

Ya se ha comprobado que el modelo está mostrando un desempeño satisfactorio, pero no se sabe aún si su población ha cambiado con el paso del tiempo. Para ello, es necesario

usar el indicador de PSI para establecer umbrales a partir de los cuales se generará una alerta. Un valor menor que 10 % indica que no existe cambio, valores mayores que 10 % y menores que 25 % indican pequeños cambios, y valores mayores que 25 % reflejan cambios significativos (Reyes Morales & Sosa Castro, 2022). En la Figura 23, se aprecia que el valor de PSI para la variable de predicción (*score*) se ha mantenido en valores cercanos a 0 %, lo que indica una estabilidad notable y se cataloga como una situación sin alerta.

**Figura 23**

*Evolutivo PSI del* score

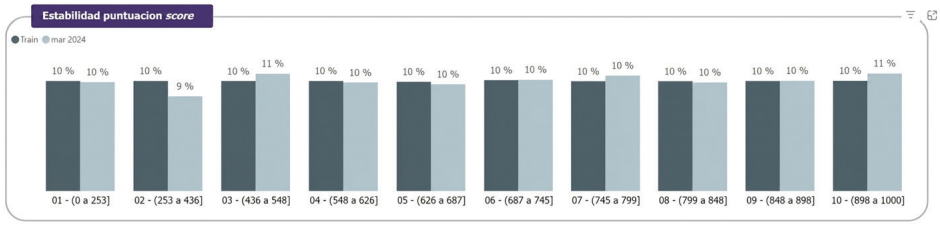



*Nota*. Diagrama extraído de la plataforma web.

La Figura 24 es consistente con lo mencionado anteriormente. Al observar la distribución, se aprecia que los valores son muy similares a los de la muestra de *train*, por lo que indica que la población no ha sufrido cambios significativos.

**Figura 24**

*Distribución del* score *vs.* train




*Nota*. Diagrama extraído de la plataforma Web.

Así como se han observado los resultados del *score*, es importante verificar la estabilidad de las variables predictoras. En la Figura 25, se observa un cuadro resumen del PSI para todas las variables, en el que ninguna de estas presenta alerta. En línea con esta idea para la variable más importante (top 1), en la Figura 26 se presenta su distribución, la cual tiene solo pequeñas variaciones con respecto a la de *train*.

**Figura 25**

*PSI por variables ordenadas de mayor a menor peso*

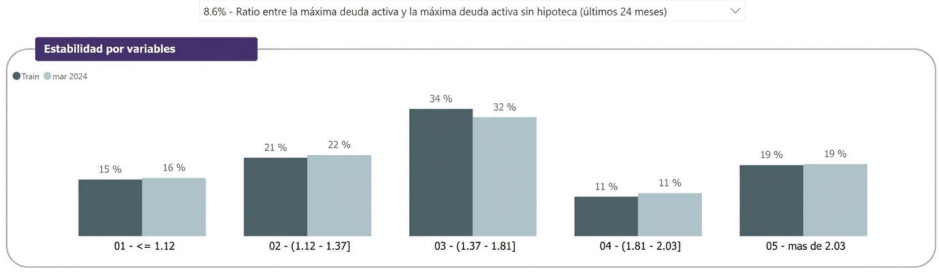
**PSI por variables**

| Variables | nov 2023 | dic 2023 | ene 2024 | feb 2024 | mar 2024 |
|---|---|---|---|---|---|
| 8,6 % - Ratio entre la máxima deuda activa y la máxima deuda activa sin hipoteca (últimos 24 meses) | 0,23 % | 0,22 % | 0,05 % | 0,27 % | 0,24 % |
| 8,6 % - Meses desde el último atraso ≥ 3 días (últimos 24 meses) | 0,22 % | 0,11 % | 0,15 % | 0,28 % | 0,73 % |
| 7,9 % - Variación en el número de entidades acreedoras (últimos 18 meses) | 0,05 % | 0,13 % | 0,19 % | 0,04 % | 0,06 % |
| 6,4 % - Máximo de entidades acreedoras (últimos 12 meses) | 0,00 % | 0,16 % | 0,05 % | 0,13 % | 0,08 % |
| 38,5 % - Máximo atraso en los (últimos 12 meses) | 0,20 % | 0,05 % | 0,28 % | 0,08 % | 0,78 % |
| 3,1 % - Meses desde el último atraso ≥ 1 día (últimos 24 meses) | 0,35 % | 0,82 % | 0,15 % | 0,29 % | 0,38 % |
| 2,9 % - Máximo número de días de atraso (últimos 24 meses) | 0,03 % | 0,06 % | 0,02 % | 0,01 % | 0,00 % |
| 2,5 % - Meses de buen comportamiento (últimos 18 meses) | 0,01 % | 0,03 % | 0,16 % | 0,09 % | 0,11 % |
| 18,9 % - Incremento de la deuda total (últimos 24 meses) | 0,01 % | 0,14 % | 0,15 % | 0,14 % | 0,04 % |
| 1,7 % - Meses desde la última clasificación distinta a 'normal' (últimos 12 meses) | 0,01 % | 0,00 % | 0,08 % | 0,00 % | 0,00 % |
| 0,7 % - Diferencia entre meses de buen y mal comportamiento (últimos 12 meses) | 0,10 % | 0,29 % | 0,07 % | 0,13 % | 0,18 % |

*Nota*. Diagrama extraído de la plataforma web.

**Figura 26**

*Distribución de la variable top 1*




*Nota*. Diagrama extraído de la plataforma web.

## DISCUSIONES

El coeficiente de Gini de 67,4 % obtenido en el periodo de seguimiento es competitivo dentro del contexto del *scoring* crediticio con datos reales de una sola institución. Estudios como el de Suhadolnik et al. (2023), en el que se trabajó con más de 2,5 millones de registros de una cartera diversificada, reportaron métricas superiores con XGBoost, lo cual es esperable dado el volumen y la variedad de datos disponibles. En cambio, este estudio se realizó con 50 206 registros de un portafolio específico, lo que restringe naturalmente el techo de rendimiento, pero, a la vez, hace que el modelo sea más representativo del comportamiento real de esa cartera. El resultado es coherente con lo reportado por De Lange et al. (2022), quienes también encontraron que los modelos desarrollados a partir de carteras reales y acotadas producen métricas más modestas pero más estables en producción.

Respecto de la explicabilidad, los valores SHAP identificaron el máximo atraso en los últimos doce meses como la variable de mayor peso, lo cual es consistente con la literatura. Castañeda Hilario et al. (2025) encontraron que la morosidad acumulada domina la predicción en carteras peruanas y De Lange et al. (2022) reportaron que el historial de comportamiento de pago es determinante en bancos europeos. El hecho de que dos contextos tan distintos converjan en el mismo tipo de variable refuerza la idea de que el modelo captura una señal genuina y no un artefacto del *dataset*.

El aspecto más diferenciador de este trabajo frente a la literatura revisada es el monitoreo automatizado. La mayoría de estudios sobre *scoring* crediticio concluyen con la validación del modelo en una muestra *out of time*, pero no abordan qué pasa después del despliegue. Bhagavathula (2025) y Shcherbinin (2026) tratan el monitoreo como problema de infraestructura genérica, mientras que este estudio lo integró directamente al ciclo de vida del modelo crediticio bajo el marco regulatorio de la SBS, lo que representa una contribución aplicada concreta.

No obstante, el estudio tiene limitaciones que deben reconocerse. La validación se realizó con datos de una sola institución financiera peruana, lo que limita la generalización de los resultados a otras carteras o entidades. El periodo de monitoreo abarca solo tres meses, un intervalo insuficiente para capturar ciclos económicos adversos o eventos de estrés como los que ocurrieron durante la pandemia. Finalmente, el modelo fue evaluado en condiciones de relativa estabilidad macroeconómica, por lo que su comportamiento ante escenarios de crisis continúa siendo una pregunta abierta para investigaciones futuras.

## CONCLUSIONES

La presente investigación logró desarrollar e implementar un sistema integral de calificación crediticia basado en el algoritmo HistGradientBoosting, optimizado mediante técnicas de búsqueda bayesiana. El modelo final alcanzó un coeficiente de Gini del 67,4 % en la etapa de seguimiento, manteniendo una diferencia mínima de un punto porcentual respecto al conjunto de entrenamiento. Este nivel de desempeño, catalogado como excelente, demuestra que la adopción de algoritmos de aprendizaje automático tipo *boosting*, combinada con una rigurosa ingeniería de características mediante *binning*, supera en robustez y precisión a los enfoques tradicionales, sin sacrificar la estabilidad temporal del riesgo calculado.

Los resultados evidenciaron que el equilibrio entre poder predictivo e interpretabilidad es alcanzable mediante el uso de valores SHAP, lo cual permitió identificar que el comportamiento histórico de pagos —específicamente, el máximo atraso en los últimos doce meses— constituye el factor determinante en la predicción del *default*. Asimismo, la implementación de un esquema de monitoreo continuo validó la estabilidad del modelo,

registrando un índice de estabilidad poblacional (PSI) de 0,004 (0,4 %) para el *score*, cifra que se ubica muy por debajo del umbral de alerta de 0,10. Esto confirma que, a pesar de la volatilidad del entorno económico, la estructura de las variables predictoras y la distribución del riesgo se mantuvieron constantes durante el periodo de evaluación.

Más allá de la precisión métrica, este estudio revela el valor práctico de integrar el ciclo de vida del modelo con una arquitectura tecnológica moderna basada en servicios en la nube (AWS) y visualización web. La capacidad de desplegar tableros de control en tiempo real para la vigilancia de métricas de *performance* (coeficiente de Gini, tasa de malos) y estabilidad (PSI) responde directamente a las exigencias regulatorias de la SBS sobre la gestión de riesgos de modelos. Esta aplicación permite a las instituciones financieras pasar de revisiones estáticas y periódicas a una gestión de riesgos dinámica y proactiva, lo que facilita la detección temprana de desviaciones o *drifts* en la cartera.

Como líneas de trabajo futuro se identifican cuatro frentes. Primero, incorporar métricas de impacto financiero, tales como la pérdida esperada y el beneficio operativo derivado de la correcta clasificación de clientes, lo que permitiría traducir el rendimiento estadístico del modelo en valor tangible para la institución. Segundo, someter el modelo a pruebas de estrés bajo escenarios macroeconómicos adversos, como crisis financieras o periodos de alta morosidad, a fin de evaluar su robustez fuera de condiciones estables. Tercero, explorar el monitoreo de variables de negocio adicionales como tasas de aprobación, rentabilidad por segmento y tasa de recuperación, para ampliar el seguimiento más allá de las métricas estadísticas. Cuarto, evaluar la incorporación de fuentes de datos no tradicionales para enriquecer el perfilamiento de clientes con historial crediticio limitado y explorar mecanismos de reentrenamiento automático activados por las alertas del sistema de monitoreo, lo que permitiría completar el ciclo de MLOps.

Finalmente, este trabajo documenta que los modelos de *machine learning* de tipo *boosting* superan a la regresión logística clásica en el contexto bancario peruano y establecen un marco metodológico replicable para su auditoría y explicabilidad. Los resultados muestran que es posible implementar algoritmos de alta complejidad cumpliendo con los principios de transparencia y gestión de riesgos exigidos por el regulador, siempre que el ciclo de vida del modelo incluya la interpretabilidad, una validación robusta y un monitoreo continuo en producción.

## REFERENCIAS

Aguirre Felix Díaz, I., Argomedo Sotelo, G. Y., Monzon Ñañez, J. A., & Tuesta Izaguirre, C. A. (2022). *Impacto de la adopción de inteligencia artificial como estrategia de negocio en las empresas del sector servicios durante la época de pandemia en el Perú* [Tesis de maestría, Pontificia Universidad Católica del Perú]. Repositorio Institucional PUCP. http://hdl.handle.net/20.500.12404/21241

Banchero, S., Verón, S., Petek, M., Sarrailhe, S., & De Abelleyra, D. (2021). *Detección de outliers en muestras de entrenamiento generadas mediante interpretación visual* [Documento de conferencia]. 50 Jornadas Argentinas de Informática e Investigación Operativa, XIII Congreso de Agroinformática (CAI 2021), Argentina. https://sedici.unlp.edu.ar/bitstream/handle/10915/140745/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Betancourt Ludeña, I. C. (2025). *Aplicación de modelos de predicción para identificar patrones de abandono escolar en los datos de unidades educativas del sistema nacional* [Tesis de maestría, Universidad de las Américas]. Repositorio Digital UDLA. https://dspace.udla.edu.ec/handle/33000/18132

Bhagavathula, M. B. (2025). ML pipeline monitor: A comprehensive OpenTelemetry—Based observability platform for production machine learning systems. *International Journal of AI, BigData, Computational and Management Studies*, *6*(2), 109-112. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I2P112

Buergo Lagunas, J. R. (2025). *Cálculo eficiente de SHAP para modelos de redes bayesianas gaussianas* [Tesis de maestría, Universidad de Cantabria]. Repositorio Abierto UCrea. https://repositorio.unican.es/xmlui/handle/10902/37755

Carbó-Valverde, S., & Rodríguez-Fernández, F. (2024). *Inteligencia artificial en banca: situación y perspectivas*. Funcas. https://www.funcas.es/libro/inteligencia-artificial-en-banca-situacion-y-perspectivas/

Castañeda Hilario, A., Mas Azahuanche, G. A., Mendoza Arenas, R. D., Castillo Paredes, O. T. A., Reinoso Palacios, A. R., Delgado Baltazar, M. P., & Mendoza Delgado, R. S. (2025). Explainable machine learning for credit card default prediction using web-scraped financial data: A case study in the Peruvian banking sector. En *Entrepreneurship with purpose: Social and Technological Innovation in the age of IA* (pp. 1-10). LACCEI. *https://laccei.org/LEIRD2025-Cartagena/meta/FP451.html*

Cepeda Cavero, L. E., & Véliz Soto, M. P. (2025). Ética y gobernanza algorítmica en el sector financiero: revisión sistemática de principios, gobernanza y brechas de implementación. *Interfases*, (22), 159-184. https://doi.org/10.26439/interfases2025.n022.8230

Cervantes Artavia, J., Monge Cordonero, M., & Sabater Guzmán, D. (2023). *Score crediticio con regresión logística y máquinas de soporte vectorial*. Universidad de Costa Rica, Escuela de Matemática. https://serengueti.fce.ucr.ac.cr/images/2023/03/16/Articulos/Score_crediticio_con_Regresion_Logistica_y_Maquinas_de_Soporte_Vectorial.pdf

Cordova Benites, L. S., Gonzales Collantes, L. Y., Leiva Cruz, F. D. R., & Navarro Espinoza, L. E. (2024). *Aplicación de la inteligencia artificial a la evaluación crediticia en el sistema financiero peruano* [Tesis de maestría, Universidad ESAN]. Repositorio Institucional ESAN. https://hdl.handle.net/20.500.12640/4249

Cosme Cervera, S., Rodríguez Arellano, S., Rocha Barrón, L. R., & Angel Angel, J. D. J. (2025, noviembre). Qué tan vulnerable es la regresión lineal ante los datos del mundo real. En D. Tinoco Varela (Ed.), *Memorias del Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología* (pp. 118-124). Universidad Nacional Autónoma de México. https://www.researchgate.net/profile/Jose-Angel-Angel/publication/397461440

Dar, U., & Cavus, M. (2026). *From XAI to MLOps: Explainable concept drift detection with profile drift detection*. arXiv. *https://doi.org/10.48550/arXiv.2412.11308*

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*, 100164. https://www.sciencedirect.com/science/article/pii/S2772662223000048

De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Dos Santos, A. B., Dos Santos, I. D. B., & Toma, K. M. (2025). Detecção de transações fraudulentas com cartão de crédito: uma abordagem comparativa baseada em aprendizado de máquina. *Revista Foco*, *18*(5), e8328. https://doi.org/10.54751/revistafoco.v18n5-037

Enriquez Montes, J., & Orbezo Huamancari, D. A. M. (2025). *Modelo de* scoring *de clientes para la gestión eficiente de productos financieros mediante regresión logística y* data analytics*: un enfoque moderno para la industria financiera en Perú* [Trabajo de suficiencia profesional, Universidad Peruana de Ciencias Aplicadas]. Repositorio Académico UPC. https://repositorioacademico.upc.edu.pe/handle/10757/684767

Erazo Laínez, G. (2024). *Machine learning para la estimación de riesgo de crédito de una cartera de consumo* [Tesis de maestría, Universidad Complutense de Madrid]. Docta Complutense. https://docta.ucm.es/entities/publication/fc512956-4b59-4f40-ab93-18850315b2a3

Ferro Rugeles, R. D., Cossio Escobar, G., & Fernández Moncada, N. (2023). *Pronóstico de ventas a través de una aplicación web aplicando machine learning y variables exógenas* [Trabajo de grado, Escuela Colombiana de Ingeniería Julio Garavito]. Repositorio Escuela Colombiana de Ingeniería Julio Garavito. https://repositorio.escuelaing.edu.co/handle/001/2815

Gantman, E. R. (2020). Apertura comercial, pobreza y desigualdad: un análisis mediante árboles de regresión. *RInERS: Revista de Investigación en Economía y Responsabilidad Social*, *1*(4), 1-14. https://repositoriocyt.unlam.edu.ar/handle/123456789/1298

García García, P. (2024). *Benchmarking de técnicas de explicabilidad para la inteligencia artificial* [Tesis de maestría, Universidad Politécnica de Madrid]. Archivo Digital UPM. https://oa.upm.es/82899/1/TFM_PABLO_GARCIA_GARCIA.pdf

Hermitaño Castro, J. A. (2022). Aplicación de *machine learning* en la gestión de riesgo de crédito financiero: una revisión sistemática. *Interfases*, (15), 160-178. https://doi.org/10.26439/interfases2022.n015.5898

Jaimes, D., González, C. A., Llanos, L., Estrada, O., Barrios, C., Agudelo, D., Navarro Racines, C., Jiménez, D., Gardeazabal, A., & Ramírez Villegas, J. (2023). *Aplicación en la agricultura de técnicas de explicabilidad del aprendizaje de máquinas: aclarando la caja negra*. Alliance Bioversity & CIAT; CGIAR; CIMMYT. https://hdl.handle.net/10568/135694

Joshi, S. (2025). *Model risk management in the era of generative AI: Challenges, opportunities, and future directions* (MPRA Paper 125221). Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/125221/

Marchant Contreras, V. M. (2022). *Un modelo predictivo interpretable para la estimación del ingreso monetario de clientes bancarios basado en XGBoost y SHAP* [Tesis de licenciatura, Universidad de Concepción]. Repositorio UdeC. https://repositorio.udec.cl/server/api/core/bitstreams/fcf68aa4-ddb0-4ef7-9cb4-661b16a7b231/content

Martínez Martínez, G. (2024). *Optimización bayesiana con multifidelidad para la búsqueda de hiperparámetros en algoritmos de aprendizaje por refuerzo* [Tesis de maestría, Universidad Politécnica de Madrida]. Archivo Digital UPM. https://oa.upm.es/82959/

Moneta Pizarro, A. M., Juárez, M. A., Camilo Caro, L. N., & Allub, M. D. H. (2022). Una revisión del supuesto MAR en datos faltantes de la EPH. *Documentos de Trabajo de Investigación de la Facultad de Ciencias Económicas (DTI-FCE)*, *7*, 1-20. https://revistas.unc.edu.ar/index.php/DTI/article/view/37746

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Preprints.org*. https://doi.org/10.20944/preprints202308.0947.v1

Pérez Monsalve, L. O. (2024). *Predicción de la vida útil de las baterías de un Nissan Leaf usando datos reales de conducción mediante el uso de Machine Learning* [Trabajo

de grado, Universidad de Los Andes]. Repositorio Institucional Séneca. https://hdl.handle.net/1992/75130

Polo Romero, V. J. (2024). Modelo predictivo basado en Naive Bayes a través de *machine learning supervised* y la deserción estudiantil, en centros de educación tecnológicos públicos de la región La Libertad. *Sciéndo Ingenium*, 59-71. https://revistas.unitru.edu.pe/index.php/PGM/article/view/6169

Putri, N. H., Fatekurohman, M., & Tirta, I. M. (2021). Credit risk analysis using support vector machines algorithm. *Journal of Physics: Conference Series*, *1863*, 012039. https://doi.org/10.1088/1742-6596/1836/1/012039

Rahmani, R., Parola, M., & Cimino, M. G. C. A. (2024). *A machine learning workflow to address credit default prediction*. arXiv. *https://doi.org/10.48550/arXiv.2403.03785*

Resolución S.B.S. 00053-2023 [Superintendencia de Bancas, Seguros y AFP]. Reglamento de Gestión de Riesgos de Modelo. 6 de enero del 2023. https://intranet2.sbs.gob.pe/dv_int_cn/2240/v1.0/Adjuntos/0053-2023.R.pdf

Reyes Morales, M. A., & Sosa Castro, M. M. (2022). Modelo de puntuación crediticia para tarjeta de crédito en México: una aproximación logística. *Ensayos Revista de Economía*, *41*(1), 17-52. https://doi.org/10.29105/ensayos41.1-2

Romero Carmen, J. A. (2024). *El margen financiero como medida de riesgo de incumplimiento de los préstamos personales del Perú entre 2019 y 2022* [Tesis de maestría, Pontificia Universidad Católica del Perú]. Repositorio Institucional PUCP. https://tesis.pucp.edu.pe/items/f0ab9c35-89a8-4b54-973a-2c1a1a8ce82e

Shcherbinin, A. (2026). Adaptive mechanisms for model retraining in production: Drift triggers, retraining prioritisation, and reduced data preparation time through automated orchestration. *Universal Library of Innovative Research and Studies*, *3*(1), 94-100. https://doi.org/10.70315/uloap.ulirs.2026.0301013

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management*, *16*(12), 496. https://doi.org/10.3390/jrfm16120496

Superintendencia de Banca, Seguros y AFP. (2025). *Oficio 02569-2025-SBS: opinión sobre el Proyecto de Ley 8969*. https://www.sbs.gob.pe/Portals/0/jer/opinion_proy_leg/2024/1-OFICIO-02569-2025-SBS-PL8969.pdf

Tocto Reyes, G., Giraldo Coral, X. A., & Hirpo Marchinares, A. E. (2022). Aplicación de *machine learning* para campañas de *marketing* en la banca comercial. *Interfases*, *2*(16), 185-198. https://www.redalyc.org/pdf/7301/730180375008.pdf

True North Partners. (2023). *PRA SS1/23 - Principios de gestión de riesgo de modelo. Nuevos requerimientos regulatorios. https://www.clubgestionriesgos.org/wp-content/uploads/TNP-Riesgo-de-Modelo-PRA-SS1-23_CGR.pdf*

Valenzuela Sánchez, M. S. (2025). *Predicción del riesgo crediticio en clientes del sector retail mediante redes neuronales: caso de estudio de la empresa Credicorp* [Tesis de maestría, Universidad de las Américas]. Repositorio Digital UDLA. https://dspace.udla.edu.ec/handle/33000/18134

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in Finance—A systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, *74*, 867-907. https://doi.org/10.1007/s11301-023-00320-0

Wheeler, R., & Carroll, F. (2023). An explainable AI solution: Exploring extended reality as a way to make artificial intelligence more transparent and trustworthy. En C. Onwubiko, P. Rosati, A. Rege, A. Erola, X. Bellekens, H. Hindy & M. G. Jaatun (Eds.), *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media* (pp. 255-276). Springer. https://link.springer.com/chapter/10.1007/978-981-19-6414-5_15

Yang, S., Huang, Z., Xiao, W., & Shen, X. (2024). *Interpretable credit default prediction with ensemble learning and SHAP*. arXiv.

Zegarra Jara, N. (Dir.). (2025). *Microfinanzas: la revista del sistema financiero peruano* (Edición 236). https://statuscomunicaciones.pe/microfinanzas/M236.pdf

Zuleta Fuerte, D. F., Bru Cordero, O. E., & Pastor Sierra, K. S. (2025). Validación cruzada: una herramienta crucial para mejorar la eficiencia de modelos de clasificación con datos biomédicos. *Comunicaciones en Estadística*, *18*(1), 51-85. https://revistas.usantotomas.edu.co/index.php/estadistica/article/view/11214